**CSE 522: Algorithms and Uncertainty** Spring 2017

Lecture 8 — April 21-24, 2017

*Lecturer: Anna R. Karlin*

# 1 Convex functions review

**Definition 1.1.** A set $S$ is convex if for every two points $\mathbf{x}, \mathbf{y} \in S$, every point on the line segment between them is also in the set. A function $f : S \subset \mathbb{R}^n \to \mathbb{R}$ is convex if for every $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0, 1]$,

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$

Some key facts about convex functions follow. For more see §7.

### Tangents lies below the function:

If $f$ is convex and differentiable,

$$\forall \mathbf{u} \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \nabla f(\mathbf{w}) \cdot (\mathbf{u} - \mathbf{w}).$$

Recall the gradient

$$\nabla f(\mathbf{w}) = \left( \frac{\partial f(\mathbf{w})}{\partial w_1}, \frac{\partial f(\mathbf{w})}{\partial w_2}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_n} \right).$$

*Proof.* $\rightarrow$:

$$f(\lambda y + (1 - \lambda)x) \leq \lambda f(y) + (1 - \lambda) f(x).$$

Therefore,

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq f(y) - f(x).$$

Taking $\lambda$ to 0, on the left hand side, we get

$$\frac{f(x) + \lambda f'(x)(y - x) - f(x)}{\lambda} \leq f(y) - f(x).$$

$\leftarrow$: Suppose that $f'(x)(y - x) \leq f(y) - f(x)$. Let

$$z = \lambda y + (1 - \lambda)x$$

. Then

$$f(y) \geq f(z) + f'(z)(y - z)$$
$$f(x) \geq f(z) + f'(z)(x - z).$$

Multiply first by $\lambda$ and second by $1 - \lambda$ and add.

To extend the proof to higher dimensions, after fixing points $x$ and $y$, it suffices to consider the line segment connecting them. □

### Subgradients

**Definition 1.2.** $\mathbf{v}$ is a **subgradient** of $f$ at $\mathbf{w}$ if for all $\mathbf{u}$,

$$f(\mathbf{u}) \geq f(\mathbf{w}) + \mathbf{v} \cdot (\mathbf{u} - \mathbf{w}).$$

The **differential set** $\partial f(\mathbf{w})$ is the set of subgradients of $f$ at $\mathbf{w}$.

**Lemma 1.3.** *$f$ is convex iff for every $\mathbf{w} \in S$, $\partial f(\mathbf{w}) \neq \emptyset$.*

*Proof.* $\leftarrow$ same as above. $\rightarrow$: see §7. □

### Local minimum = global minimum

Most important property and the thing that makes convex functions so nice and relatively easy to optimize

**Lemma 1.4.** *Let $f$ be a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$, and let $\mathbf{x} \in S$. Then the following are equivalent:*

1. *$\mathbf{x}$ is a global minimum*
2. *$\mathbf{x}$ is a local minimum*
3. *$\nabla f(\mathbf{x}) = 0$.*

*(a) to (b) immediate. (b) to (c) holds for any function. (c) to (a) follows from*

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) = f(x).$$

More generally, if $S$ be a closed convex subset of $\mathbb{R}^n$ and $f : S \to \mathbb{R}$ is a convex, differentiable function. Then $\mathbf{x} \in S$ is a global minimum iff

$$(\nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}) \geq 0$$

for all $\mathbf{y} \in S$.

## 2 Online convex optimization

- Play $\mathbf{w}_t$ from a convex and compact subset $S$ of a linear space.
- Observe the convex loss $\ell_t : S \to \mathbb{R}$, and pay $\ell_t(\mathbf{w}_t)$.
- Update $\mathbf{w}_t \to \mathbf{w}_{t+1} \in S$.

**Definition 2.1.** The **regret** of the online convex optimization algorithm:

$$\text{Regret}_T(\mathbf{u}) = \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u})$$

and

$$\text{Regret}_T = \max_{\mathbf{u} \in S} \text{Regret}_T(\mathbf{u}).$$

# 3 Finding a good online algorithm for minimizing regret

## 3.1 Follow the leader

The most obvious approach is to "follow the leader".

**Definition 3.1.** The **Follow the Leader (FTL)** strategy is to set

$$\mathbf{w}_{t+1} = \mathsf{argmin}_{\mathbf{w}\in S} \sum_{s=1}^{t} \ell_s(\mathbf{w}).$$

The key lemma that we will use is the following:

**Lemma 3.2.** *For all* $\mathbf{u} \in S$*,*

$$\text{Regret}_T(\mathbf{u}) = \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u}) \leq \sum_{t=1}^{T} (\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1})).$$

*Remark* 3.3. The lemma is useful because it shows that if we can guarantee that $\ell_t(\mathbf{w}_t)$ is generally close to $\ell_t(\mathbf{w}_{t+1})$ for all $t$, then FTL will in fact work well. In other words, we want the predictions to be stable. This lemma shows that as long as the predictions are "stable", we will have low regret. This is exactly what Nikhil did when he was analyzing FTPL.

*Proof.* The statement that for all $\mathbf{u}$

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u}) \leq \sum_{t=1}^{T} (\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}))$$

is the same as

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_{t+1}) \leq \sum_{t=1}^{T} \ell_t(\mathbf{u}).$$

This is precisely the statement that "Be the leader" is as good or better than any fixed strategy. See proof below. □

## 3.2 Be The Leader (BTL)

At each time $t$, set $\mathbf{w}_t := \mathsf{argmin}_{w\in S} \sum_{s=1}^{t} \ell_s(\mathbf{w})$.

Notice that this algorithm is not implementable because it requires that the algorithm know $\ell_t(\cdot)$.

**Theorem 3.4.** BTL *is at least as good as the best fixed strategy, that is, for all* $\mathbf{u}$*,*

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) \leq \sum_{t=1}^{T} \ell_t(\mathbf{u}).$$

*Proof.* Proof by induction on $T$. Base case follows from definition. Assume inequality holds for $T-1$. Then for all $\mathbf{u} \in S$, we have

$$\sum_{t=1}^{T-1} \ell_t(\mathbf{w}_t) \leq \sum_{t=1}^{T-1} \ell_t(\mathbf{u}).$$

Adding $\ell_T(\mathbf{w}_T)$ to both sides, we get

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) \leq \ell_T(\mathbf{w}_T) + \sum_{t=1}^{T-1} \ell_t(\mathbf{u}).$$

Since this holds for all $\mathbf{u}$, in particular it holds for $\mathbf{u} := \mathbf{w}_T$. Thus,

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) \leq \sum_{t=1}^{T-1} \ell_t(\mathbf{w}_T) = \min_{\mathbf{u} \in S} \sum_{t=1}^{T} \ell_t(\mathbf{u}).$$

□

# 4 Online quadratic optimization

For online quadratic optimization, FTL does produce stable predictions.

Suppose that $\ell_t(\mathbf{w}) = \frac{1}{2}||\mathbf{w} - \mathbf{z}_t||_2^2$ in each round for some vector $\mathbf{z}_t$.

For this problem, FTL chooses

$$\mathbf{w}_{t+1} = \mathsf{argmin}_{\mathbf{w}} \sum_{1 \leq i \leq t} \frac{1}{2}||\mathbf{w} - \mathbf{z}_i||_2^2.$$

Looking at the gradient, we have

$$\sum_{1 \leq i \leq t} (\mathbf{w}_{t+1} - \mathbf{z}_i) = 0$$

or equivalently

$$\mathbf{w}_{t+1} = \frac{1}{t} \sum_{i=1}^{t} \mathbf{z}_i,$$

the average of the $\mathbf{z}$'s. Observe that

$$\mathbf{w}_{t+1} = \frac{1}{t}(\mathbf{z}_t + (t-1)\mathbf{w}_t) = \left(1 - \frac{1}{t}\right)\mathbf{w}_t + \frac{1}{t}\mathbf{z}_t,$$

which yields

$$\mathbf{w}_{t+1} - \mathbf{z}_t = \left(1 - \frac{1}{t}\right)(\mathbf{w}_t - \mathbf{z}_t).$$

Therefore

$$\begin{aligned}
\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1}) &= \frac{1}{2}||\mathbf{w}_t - \mathbf{z}_t||^2 - \frac{1}{2}||\mathbf{w}_{t+1} - \mathbf{z}_{t+1}||^2 \\
&= \frac{1}{2}\left(1 - \left(1 - \frac{1}{t}\right)^2\right)||\mathbf{w}_t - \mathbf{z}_t||^2 \\
&\leq \frac{1}{t}||\mathbf{w}_t - \mathbf{z}_t||^2.
\end{aligned}$$

If

$$L = \max_t ||\mathbf{z}_t||$$

then since $\mathbf{w}_t$ is an average of the $\mathbf{z}$'s, by the triangle inequality

$$||\mathbf{w}_t - \mathbf{z}_t|| \leq 2L.$$

Therefore,

$$\sum_t [\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1})] \leq 4L^2 \sum_{t=1}^{T} \frac{1}{t} \leq 4L^2(\ln T + 1).$$

Combining this with the BTL lemma which says that

$$R_T(\mathbf{u}) = \sum_{t=1}^{T} [\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})] \leq \sum_{t=1}^{T} [\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}_{t+1})] = O(L^2 \log T).$$

### 4.1 No good for linear losses

Regret can be linear due to **l**ack of stability. Suppose that $S = [-1, +1]$, and $\ell_1(w) = w/2$. Then if

$$\ell_t(w) = \begin{cases} -w & \text{if } t \text{ is even} \\ +w & \text{if } t \text{ is odd} \end{cases}$$

This implies that

$$\sum_{s=1}^{t} \ell_s(w) = \begin{cases} -\frac{w}{2} & \text{if } t \text{ is even} \\ +\frac{w}{2} & \text{if } t \text{ is odd} \end{cases}$$

Therefore after even steps, the leader is $w = 1$ and after odd steps, the leader is $w = -1$. Which means that on odd steps, the loss will be 1 and on even steps the loss will be 1. On the other hand, if left $w = 0$ every step, the loss would be 0. So the regret is linear.

## 5 Online gradient descent [Zinkevich '03]

The most basic approach to online convex optimization comes from offline optimization, and that is based on the following: If you want to minimize a convex function, move in the direction of steepest descent, For a convex function this is guaranteed to converge to global minimum.

Suggests the following idea in the online setting:

**Definition 5.1.** The **Online Gradient Descent (OGD)** strategy consists of repeatedly taking a small step in the direction of steepest descent (negative gradient) and then, if you're outside the convex set $S$, projecting back to closest point in $S$.

After step $t$, given $\ell_t(\cdot)$:

$$\mathbf{y}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell_t(\mathbf{w}_t)$$

$$\mathbf{w}_{t+1} = \mathsf{argmin}_{\mathbf{w} \in S} ||\mathbf{y}_{t+1} - \mathbf{w}||$$

**Theorem 5.2.** *OGD has regret $O(\sqrt{T})$.*

*Proof.* Using convexity, we can bound

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) \leq \nabla \ell_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{u}) = \nabla_t \cdot (\mathbf{w}_t - \mathbf{u}) \tag{1}$$

where

$$\nabla_t := \nabla \ell_t(\mathbf{w}_t).$$

To bound $\mathbf{w}_t - \mathbf{u}$, first observe that it follows from Pythagoras' Theorem (see below) that

$$||\mathbf{w}_{t+1} - \mathbf{u}|| \leq ||\mathbf{y}_{t+1} - \mathbf{u}||.$$

In turn, by definition

$$||\mathbf{y}_{t+1} - \mathbf{u}||^2 = ||\mathbf{w}_t - \eta_t \nabla_t - \mathbf{u}||^2 = ||\mathbf{w}_t - \mathbf{u}||^2 - 2\eta_t \nabla_t \cdot (\mathbf{w}_t - \mathbf{u}) + \eta_t^2 ||\nabla_t||^2.$$

Combining this with the previous inequality, we get that

$$2\eta_t \nabla_t \cdot (\mathbf{w}_t - \mathbf{u}) \leq ||\mathbf{w}_t - \mathbf{u}||^2 - ||\mathbf{w}_{t+1} - \mathbf{u}||^2 + \eta_t^2 ||\nabla_t||^2. \tag{2}$$

Combining (1) and (2), we have

$$\begin{aligned}
\sum_{t=1}^{T} [\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u})] &\leq \sum_{t=1}^{T} \nabla_t \cdot (\mathbf{w}_t - \mathbf{u}) \\
&\leq \sum_{t=1}^{T} \frac{1}{2\eta_t} \left( ||\mathbf{w}_t - \mathbf{u}||^2 - ||\mathbf{w}_{t+1} - \mathbf{u}||^2 + \eta_t^2 ||\nabla_t||^2 \right) \\
&= \sum_{t=1}^{T} ||\mathbf{w}_t - \mathbf{u}||^2 \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) + \sum_{t=1}^{T} \frac{\eta_t}{2} ||\nabla_t||^2 \qquad \text{where } \frac{1}{\eta_0} := 0 \\
&\leq \frac{D^2}{2} \frac{1}{\eta_T} + \frac{G^2}{2} \sum_{t=1}^{T} \eta_t,
\end{aligned}$$

where the last line follows from the substitutions

$$D = \max_{\mathbf{x} \in S} ||\mathbf{x} - \mathbf{u}|| \quad \text{and} \quad G = \max_t ||\nabla_t||,$$

and the telescoping series. Finally, setting

$$\eta_t := \frac{D}{G\sqrt{2t}}$$

we obtain the following upper bound on the regret

$$R_T(\mathbf{u}) \leq \sqrt{2} D G \sqrt{T}.$$

□

**Claim 5.3.** *Let $\mathbf{u} \in S$, where $S$ is a convex set, and let*

$$\mathbf{w} = \mathsf{argmin}_{\mathbf{x} \in S} ||\mathbf{y} - \mathbf{x}||.$$

*Then*

$$||\mathbf{w} - \mathbf{u}|| \leq ||\mathbf{y} - \mathbf{u}||$$

.

*Proof.* Wlog assume that $\mathbf{y} = \mathbf{0}$. Thus, we need to show that $||\mathbf{w} - \mathbf{u}|| \leq ||$
$bu||$. This follows from

$$\begin{aligned} ||\mathbf{w}||^2 \leq ||(1-\epsilon)\mathbf{w} + \epsilon\mathbf{u}||^2 &= ||(\mathbf{w} + \epsilon(\mathbf{u} - \mathbf{w})||^2 \\ &= ||\mathbf{w}||^2 + \epsilon^2||\mathbf{u} - \mathbf{w}||^2 + 2\epsilon(\mathbf{w}, \mathbf{u} - \mathbf{w}) \end{aligned}$$

Taking $\epsilon$ to 0 implies that $(\mathbf{w}, \mathbf{u} - \mathbf{w}) \geq 0$. Thus, $||\mathbf{w}||^2 \leq (\mathbf{u}, \mathbf{w})$, so

$$||||\mathbf{w} - \mathbf{u}||^2 = ||\mathbf{w}||^2 - 2(\mathbf{u}, \mathbf{w}) + ||\mathbf{u}||^2 \leq ||\mathbf{u}||^2 - (\mathbf{u}, \mathbf{w}) \leq ||\mathbf{u}||^2.$$

□

## Lower bound

Any algorithm for online convex optimization incurs $\Omega(DG\sqrt{T})$ regret in the worst case, even if the loss functions are generated from a fixed stationary distribution.

Specifically, suppose that OCO is over $n$-dimensional hypercube,

$$S = \{\mathbf{x} \in \mathbb{R}^n \mid ||\mathbf{x}||_\infty \leq 1\},$$

and there are $2^n$ linear cost functions, one associated with each vertex in $\{\pm 1\}^n$. Specifically, for each $\mathbf{v} \in \{\pm 1\}^n$, there is an associated loss function

$$\ell_{\mathbf{v}}(\mathbf{x}) = \mathbf{v} \cdot \mathbf{x},$$

and suppose that in each step the loss function is selected u.a.r. Then any algorithm incurs regret $\Omega(DG\sqrt{T})$.

## OGD for strongly convex loss functions

If the loss functions are $\alpha$-strongly convex (we'll see the definition shortly), then OGD with step-sizes $\eta_t = \frac{1}{\alpha t}$ has regret $O(G^2 \log T/\alpha)$.

# 6 Application: Stochastic Gradient Descent

Suppose that we are trying to solve an offline optimi zation problem, that is, to find the minimum of a convex function $f$ over a convex set. However, rather than being able to compute gradients, we are only able to get an approximate gradient whose expected value is correct. That is, for the function $f$ of interest, and any $\mathbf{x}$ we will have access to

$$\tilde{\nabla}_{\mathbf{x}} \quad \text{s.t.} \quad \mathbb{E}\left[\tilde{\nabla}_{\mathbf{x}}\right] = \nabla f(\mathbf{x}), \text{ and } \mathbb{E}\left[||\tilde{\nabla}_{\mathbf{x}}||^2\right] \leq G^2.$$

We can use $T$ iterations of the OGD algorithm we just looked at to find a point whose expected value is at most $O(T^{-1/2})$ larger than the true value of the minimum.

**Definition 6.1.** The **Stochastic Gradient Descent (SGD)** algorithm:

Inputs: $f$, function to be minimized,
$S$, convex set minimizing over,
$\mathbf{x}_1 \in S$, step sizes $\eta_t$

for $t := 1$ to $T$
$\mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t \tilde{\nabla}_{\mathbf{x}_t}$
$\mathbf{x}_{t+1} = \mathsf{argmin}_{\mathbf{x}\in S}||\mathbf{y}_{t+1} - \mathbf{x}||$

Return: $\mathbf{w} := \dfrac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t.$

**Theorem 6.2.** *With step sizes $\eta_t = D/(G\sqrt{t})$, SGD run for $T$ steps guarantees that*

$$\mathbb{E}\left[f(\mathbf{w})\right] \leq \min_{\mathbf{x}\in K} f(\mathbf{x}) + \frac{3GD}{2\sqrt{T}}.$$

*Proof.* Suppose that $\mathbf{x}^* = \mathsf{argmin}_x f(x)$.

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{w})\right] - f(\mathbf{x}^*) \\
&= \mathbb{E}\left[f(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t)\right] - f(\mathbf{x}^*) \\
&\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}_t)\right] - f(\mathbf{x}^*) && \text{convexity of } f \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right] \\
&\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\nabla f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}^*)\right] && \text{convexity of } f \\
&\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\tilde{\nabla}_{\mathbf{x}_t}(\mathbf{x}_t - \mathbf{x}^*)\right] && \text{noisy estimator}
\end{aligned}$$

The final inequality follows by observing that

$$\mathbb{E}\left[\tilde{\nabla}_{\mathbf{x}_t}(\mathbf{x}_t - \mathbf{x}^*)|\mathbf{x}_{t-1}, \tilde{\nabla}_{\mathbf{x}_{t-1}}\right] = \mathbb{E}\left[\tilde{\nabla}_{\mathbf{x}_t}|\mathbf{x}_{t-1}, \tilde{\nabla}_{\mathbf{x}_{t-1}}\right](\mathbf{x}_t - \mathbf{x}^*) = \nabla f(\mathbf{x}_t)\cdot(\mathbf{x}_t - \mathbf{x}^*).$$

followed by taking an outer expectation. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{w})\right] - f(\mathbf{x}^*) \\
&\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*))\right] && \text{defining } f_t(x) := \tilde{\nabla}_{\mathbf{x}_t}\cdot\mathbf{x} \\
&\leq \frac{D^2}{2}\frac{1}{\eta_T} + \frac{1}{2}\sum_{t=1}^{T}\eta_t\mathbb{E}\left[||\tilde{\nabla}_{\mathbf{x}_t}||^2\right], && \text{Algorithm is OGD on } f_t\text{'s.} \\
&\leq \frac{\sqrt{2}GD}{\sqrt{T}}.
\end{aligned}$$

Note that we were using the fact that the regret bounds hold against an adaptive adversary. □

### 6.1 Typical use of SGD

Often, SGD is used is as a substitute for regular old offline gradient descent. Suppose that you have a bunch of training examples, $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_m, b_m)$, and your goal is to minimize a function of the form

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m}\ell_{(\mathbf{a}_i, b_i)}(\mathbf{x}).$$

Then one way to generate an unbiased estimator for the gradient is to set

$$\tilde{\nabla}_{\mathbf{x}} = \nabla\ell_{(\mathbf{a}_i, b_i)}(\mathbf{x})$$

where $(\mathbf{a}_i, b_i)$ is chosen uniformly at random from the examples. Computing the gradient of a single random example is much cheaper than computing the gradient of the entire function. This returns an $\epsilon$-approximate solution after $T = O(\epsilon^{-2})$ iterations.

This matches the convergence rate of standard offline gradient descent, but each iteration is much cheaper since only one example from the data set is considered. On the other hand, the guarantee is only on the expectation.

# 7 Convex functions

We review some basic facts about convex functions:

1. A function $f : [a, b] \to \mathbb{R}$ is convex if for all $x, z \in [a, b]$ and $\alpha \in (0, 1)$ we have

$$f(\alpha x + (1 - \alpha)z) \leq \alpha f(x) + (1 - \alpha)f(z). \tag{3}$$

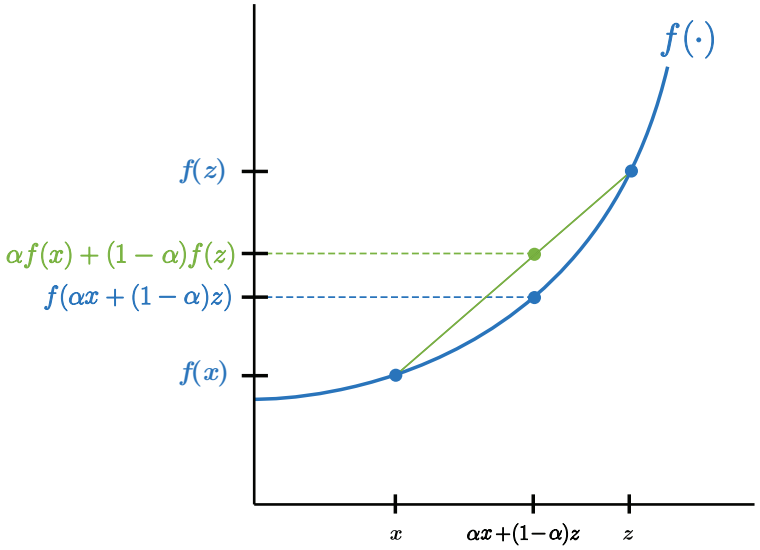


Figure 1: A convex function $f$.

2. The definition implies that the supremum of any family of convex functions is convex.

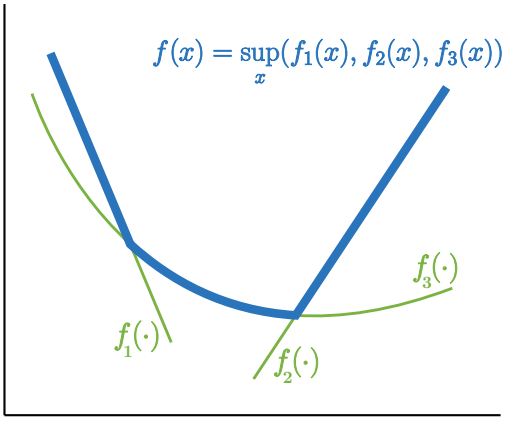


Figure 2: The supremum of three convex functions.

3. For $x < y$ in $[a, b]$ denote by $S(x, y) = \frac{f(y)-f(x)}{y-x}$ the slope of $f$ on $[x, y]$. Convexity of $f$ is equivalent to the inequality

$$S(x, y) \leq S(y, z)$$

holding for all $x < y < z$ in $[a, b]$.

4. For $x < y < z$, the inequality in (3) is equivalent to $S(x, y) \leq S(x, z)$ and to $S(x, z) \leq S(y, z)$. Thus, for $f$ convex in $[a, b]$, the slope $S(x, y)$ is (weakly) monotone increasing in $x$ and in $y$ as long as $x, y$ are in $[a, b]$. This implies continuity of $f$ in $(a, b)$.

5. It follows from (3) and the Mean Value Theorem that if $f$ is continuous in $[a, b]$ and has a (weakly) increasing derivative in $(a, b)$, then $f$ is convex in $[a, b]$.

6. The monotonicity in (4) implies that a convex function $f$ in $[a, b]$ has an increasing right derivative $f'_+$ in $[a, b)$ and an increasing left derivative $f'_-$ in $(a, b]$. Since $f'_+(x) \leq f'_-(y)$ for any $x < y$, we infer that $f$ is differentiable at every point of continuity in $(a, b)$ of $f'_+$.

7. Since increasing functions can have only countably many discontinuities, a convex function is differentiable with at most countably many exceptions. The convex function $f(x) = \sum_{n\geq 1} |x - 1/n|/n^2$ indeed has countably many points of nondifferentiability.

8. *Definition:* We say that $s \in \mathbb{R}$ is a **subgradient** of $f$ at $x$ if

$$f(y) \geq f(x) + s \cdot (y - x) \qquad \forall y \in [a, b]. \tag{4}$$

The right-hand side as a function of $y$ is called a **supporting line** of $f$ at $x$. See Figure 3 and Figure 4

9. If $s(t)$ is a subgradient of $f$ at $t$ for each $t \in [a, b]$, then

$$s(y)(x - y) \leq f(x) - f(y) \leq s(x)(x - y) \quad \forall x, y \in [a, b]. \tag{5}$$

These inequalities imply that $s(\cdot)$ is weakly increasing and $f(\cdot)$ is continuous on $[a, b]$.

10. *Fact:* Let $f : [a, b] \to \mathbb{R}$ be any function. Then $f$ has a subgradient for all $x \in [a, b]$ if and only if $f$ is convex and continuous on $[a, b]$.

*Proof:*

$\Longrightarrow$: $f$ is the supremum of affine functions (the supporting lines). Continuity at the endpoints follows from the existence of a subgradient at these points.

$\Longleftarrow$: By (4), any $s \in [f'_-(x), f'_+(x)]$ is a subgradient.

11. *Proposition:* If $s(x)$ is a subgradient of $f$ at $x$ for every $x \in [a, b]$, then

$$f(t) = f(a) + \int_a^t s(x)\, dx \qquad \forall t \in [a, b].$$

*Proof:* By translation, we may assume that $a = 0$. Fix $t \in (0, b]$ and $n > 1$. Define

$$t_k := \frac{kt}{n}.$$

For $x \in [t_{k-1}, t_k)$ , define

$$g_n(x) = s(t_{k-1}) \quad \text{and} \quad h_n(x) = s(t_k).$$

Then $g_n(\cdot) \le s(\cdot) \le h_n(\cdot)$ in $[0, t)$, so

$$\int_0^t g_n(x)\,dx \le \int_0^t s(x)\,dx \le \int_0^t h_n(x)\,dx. \tag{6}$$

By (5),

$$\frac{t}{n} s(t_{k-1}) \le f(t_k) - f(t_{k-1}) \le \frac{t}{n} s(t_k) \quad \forall k \in [1, n].$$

Summing over $k \in [1, n]$ yields

$$\int_0^t g_n(x)\,dx \le f(t) - f(0) \le \int_0^t h_n(x)\,dx. \tag{7}$$

Direct calculation gives that

$$\int_0^t h_n(x)\,dx - \int_0^t g_n(x)\,dx = [s(t) - s(0)]\frac{t}{n},$$

so by (6) and (7), we deduce that

$$\left| f(t) - f(0) - \int_0^t s(x)\,dx \right| \le [s(t) - s(0)]\frac{t}{n}.$$

Taking $n \to \infty$ completes the proof.

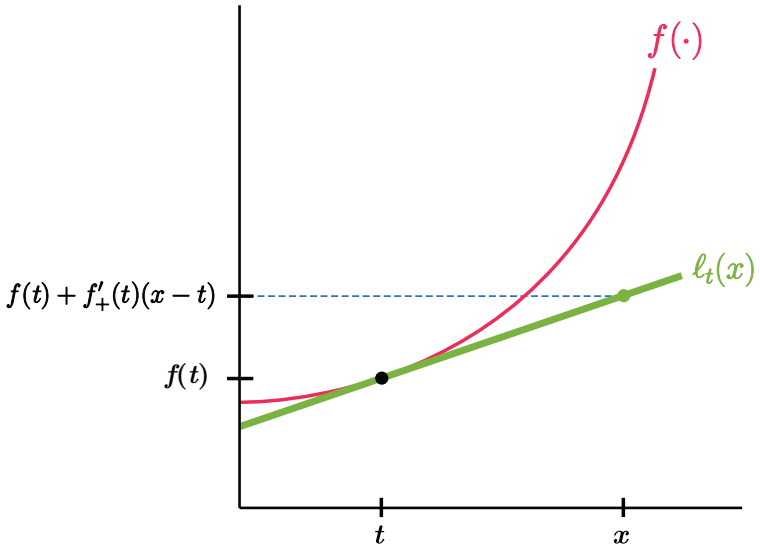


Figure 3: The line $\ell_t(\cdot)$ is a supporting line at $t$ and $f'_+(t)$ is a subgradient of $f$ at $t$.

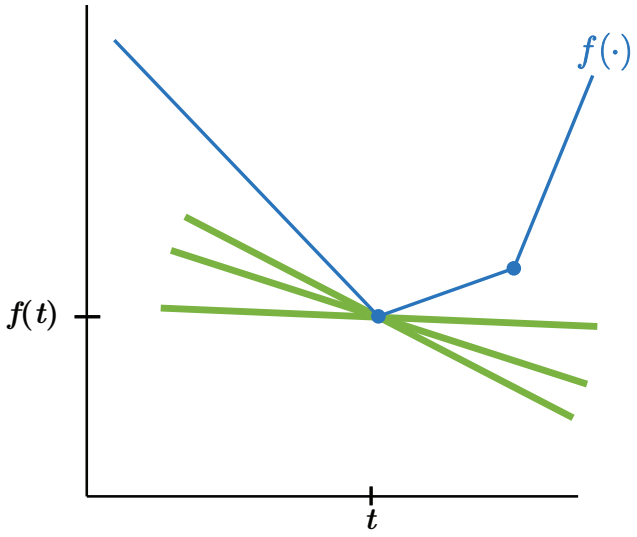


Figure 4: A collection of supporting lines at $t$.

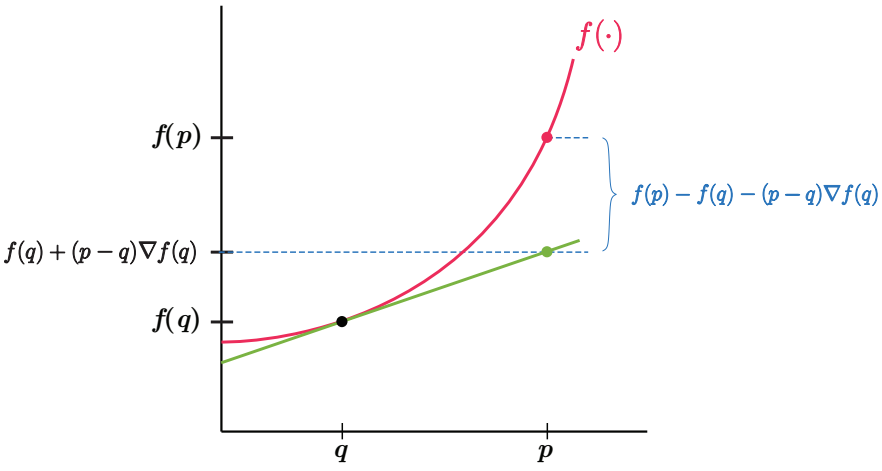


Figure 5:

12. **Jensen's inequality:** If $f : [a, b] \to \mathbb{R}$ is convex and $X$ is a random variable taking values in $[a, b]$, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. (Note that for $X$ taking just two values, this is the definition of convexity.)

    *Proof:* Let $\ell(\cdot)$ be a supporting line for $f$ at $\mathbb{E}[X]$. Then by linearity of expectation,

    $$f(\mathbb{E}[X]) = \ell(\mathbb{E}[X]) = \mathbb{E}[\ell(X)] \leq \mathbb{E}[f(X)].$$

13. The definition (3) of convex functions extends naturally to any function defined on a convex set $K$ in a vector space. Observe that the function $f : K \to R$ is convex if and only if for any $\mathbf{x}, \mathbf{y} \in K$, the function

    $$\Psi(t) = f(t\mathbf{x} + (1-t)\mathbf{y})$$

    is convex on $[0, 1]$. It follows that

    $$\Psi(1) \geq \Psi(0) + \Psi'_+(0);$$

    i.e., for all $\mathbf{x}, \mathbf{y} \in K$,

    $$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}). \tag{8}$$

    A vector $\mathbf{v} \in \mathbb{R}^n$ is a **subgradient** of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{y}$ if for all $\mathbf{x}$

    $$f(\mathbf{x}) \geq f(\mathbf{y}) + \mathbf{v} \cdot (\mathbf{x} - \mathbf{y}).$$

    If $f$ is differentiable at $\mathbf{y}$, then the only subgradient is $\nabla f(\mathbf{y})$.